Pattern Based Document Clustering and Classification in Text Corpus

Seema V. Kharate
M.E.Computer Enginnering Department,Matoshree college of Engineering and Research Center, Maharashtra, India

ABSTRACT

The big issue of extracting the feature from document in text mining because of the large number of terms, phrases, and noise. The existing methods for text mining are based on term-based approaches which extract terms from a training set for describing relevant information. In this project, novel pattern discovery approach for text mining is proposed. This approach discovers closed sequential patterns in text documents for identifying the most informational contents of the documents and then utilize the these informative contents to extract useful features for text mining. To evaluate the proposed approach, need to adopt the feature extraction method for information filtering. The expected outcomes may tested on RCV1 and TREC , the proposed approach may achieve excellent performance.

Keyword : - *Text mining, feature extraction, RCV1, TREC.*

1. INTRODUCTION

Relevant feature selection is general concept from text dimensionality of data for effective text categorization. In text classification Feature selection mainly focuses on identifying relevant information without affecting the accuracy of the classifier. The main objective of the feature selection to find the perspective patterns in the documents. The newly extracted feature set contains fewer features than the original feature set. It produces good results over a reduced dimension feature space. The issue is how to use the patterns to weight the feature accurately. This paper proposes an innovative technique for finding and classifying low-level terms based on both their presence in the patterns and their specificity in a training set document and it is used in knowledge discovery and information retrieval. Text mining refers to the discovery of useful knowledge in text documents. It is big issue to find exact required knowledge in text documents to provide data as per perspective need. The knowledge extracted from the large amount of data is beneficial in many application, such as market analysis and business management. The discovered pattern can be updated effectively and knowledge discovery also use efficiently and apply it to the field of text mining[17][19]. Data mining is therefore an essential step in the process of knowledge discovery in databases, which means data mining is having all methods of knowledge discovery process and also has modeling phase which is application of methods and algorithm for prediction of search pattern or models. Text mining is the technique that helps users find useful information from a huge digital text data[16]. It is therefore crucial that a good text mining model should fetch the information that users require with significant efficiency. Traditional Information Retrieval (IR) has the same objective of significantly retrieving as many relevant documents as possible during filtering out irrelevant documents at the same time. However, IR-based systems do not appropriately provide users with what they really need. Many text mining methods have been developed in order to achieve the goal of retrieving for information for users. We focus on the advancement of a knowledge discovery model to adequately use and update the discovered patterns and apply it to the field of text mining. The measure of knowledge discovery may subsist as following: Data Selection, Data Processing, Data Transaction, Pattern Discovery , Pattern Evaluation[21].

Most of the feature selection methods are based on term frequency or document frequency in text category. Term frequency refers to the number of time that word occurs in particular document and document frequency counts appearance of the word in number of .document . Text mining requires per-processing which the text must be disintegrate into smaller units such as terms and phrases. For example, in some text mining operation, terms extracted from the documents and treated as features. Text clustering is also referred as document clustering. Clustering is used to group the documents into relevant topics. Each of that group is refer as clusters. This is an separately learning technique. The main issue in document clustering is its high dimensionality. It requires useful algorithms which can solve this high dimensionality clustering. Several algorithms are used for text clustering which

consist of separating clustering algorithm, Density-based clustering algorithm, Model-based clustering algorithm, rank clustering algorithm and frequent pattern-based clustering. The high measurement of data is the great challenge for effective text categorization is high dimensionality. Each document in a document quantity include much riotous and irrelevant information which may reduces the efficiency for text categorization[18]. Most of the organization techniques reduce this features by eliminating stemming or stop words. It is necessary to use feature selection mechanism to hold the high capacity of data for useful text organization. In text organization Feature selection mainly focuses on finding relevant information without affecting the accuracy of the classifier. The goal of the feature selection to find the effective patterns in the documents. Feature reduction will convert the primitive features into new features by applying some conversion function. This new feature set contains less number of features or dimensions than the original set. It produces best results over a reduced dimension feature space. The challenging problem is how to use the patterns to weight accurately.

2. RELATED WORK

A literature review is an evaluation of the information found in the literature related to a particular area of study. In order to utilize these benefits, this literature review examines previous research on related topic.

2.1Text Mining

Text mining is data mining, as the application of algorithm as well as methods from the machine learning and statistics to text with goal of searching perspective pattern, Whereas data mining belongs in the corporate world because that's where more databases are, text mining assures to move machine learning technology out of the companies and into the home" as growing necessary Internet adjunct (Witten and Frank, 2000) i.e., as "web data mining" (Hearst, 1997). Laender, Ribeiro-Neto, da Silva(2001) provide a current review of web data extraction tools. Text mining is also known as text data mining, roughly parallel to text analytics, it refers to process of deriving high quality information produces text. and high quality of information is extracted through devising of patterns[21]. Text analysis imply information retrieval, lexical analysis, word frequency spreading, pattern recognition, information extraction, and data mining techniques consist of link and association analysis, visualization to turn text into data for analysis via..NLP and an alytical methods. Other way we called -Text mining is a variation on field called data mining, that tries to find perspective patterns from huge datasets. This is a concept of text mining describe in this section.

2.2 Pattern Discovery

The pattern used as a word or phrase that is extracted from the text document. There are numbers of patterns which may be identified from a text document, but not all of them are interesting. Only those evaluated to be interesting in some manner are viewed as useful knowledge. It is mid field task between association rule mining and inductive learning. It goal of finding patterns in labeled data that are descriptive[21]. A system may encounter a problem where a identified pattern is not interesting a user. Such patterns are not qualified as knowledge. Therefore, a knowledge discovery system should have the capacity of deciding whether a pattern is interesting enough to form knowledge in the current context.

2.3 Pattern Taxonomy

Pattern can be framed into taxonomy-used knowledge discovery model is developed towards applying data mining approach to practical text mining operations. Knowledge Discovery in Databases (KDD) can be referred to as the term of data mining which aims for identifying interesting patterns or trends from a database. In particular, a process of turning low-level data into high-level knowledge is indicated as KDD. The concept of KDD process is the data mining for extracting patterns from data[21]. we focus on development of knowledge discovery model to effectively use and update discovered patterns and apply it to the field of text mining.

2.4 Deploying Method

In this section, we develop equations for deploying patterns over low-level terms by evaluating term supports based on their presence in patterns. The evaluation of term supports (weights) in this paper is different from term-based approaches. For this approach, the evaluation of a given terms weight is based on its appearances in documents. In this research, terms are scored according to their appearances in discovered patterns. To improve the efficiency of the pattern taxonomy mining , an algorithm, SPMining, was proposed to find closed sequential patterns for all documents , which used the well-known Apriori property can reduce the searching space. For all positive documents, the SPMining algorithm can identify all closed sequential patterns.

2.5 Data

We used two most suitable data sets to test the proposed model: Reuters Corpus Volume 1, a very large data collection; and Reuters-21578, a small one. RCV1 has 806,791 documents that cover a broad spectrum of issues or topics. TREC (2002) has developed and provided 50 reliable assessor topics for RCV1, aiming at testing robust information filtering systems. These topics were estimated by human assessors at the National Institute of Standards and Technology (NIST)[1]. For each topic, a subset of RCV1 documents is further divided into a training set and a testing set. RCV1 is a standard data collection and the TREC 50 topics are reliable and acceptable enough for high quality experiments. Reuters-21578 corpus is generally used collection for text mining. In this experiment, we picked up the set of 10 classes. According to Sebastian is convention, it was also called R8 because two classes corn and wheat are intimately related to the class grain, and they were attached to class grain.

3. PROBLEM STATEMENT

To design a system for extracting the relevant features from text document according to a)The features provided like paragraph follow title, b) the paragraph location in document, c) Appearance of Sentence in paragraph, and d) sentence length or number of thematic word by selecting single feature or all feature at a time. The basic idea has been deployed from [1] and an extension to the system is feature to find the similarity with more accuracy than the existing algorithm and Email Author Identification.

4. IMPLEMENTATION

4.1 System Architecture

To design a system for extracting the relevant features from text document according to a)The features provided like paragraph follow title, b) the paragraph location in document, c) Appearance of Sentence in paragraph ,and d) sentence length or number of thematic word by selecting single feature or all feature at a time. The basic idea has been deployed from [1] and an extension to the system is feature to find the similarity with more accuracy than the existing algorithm and Email Author Identification. 1.Document corpus- the user will choose the set of documents for checking a relevant features and similar documents within the corpus. System scans the documents uploaded by the user. 2.Preprocessing of Document- the set of text document are uploaded for the processing which removes stop words and special characters.

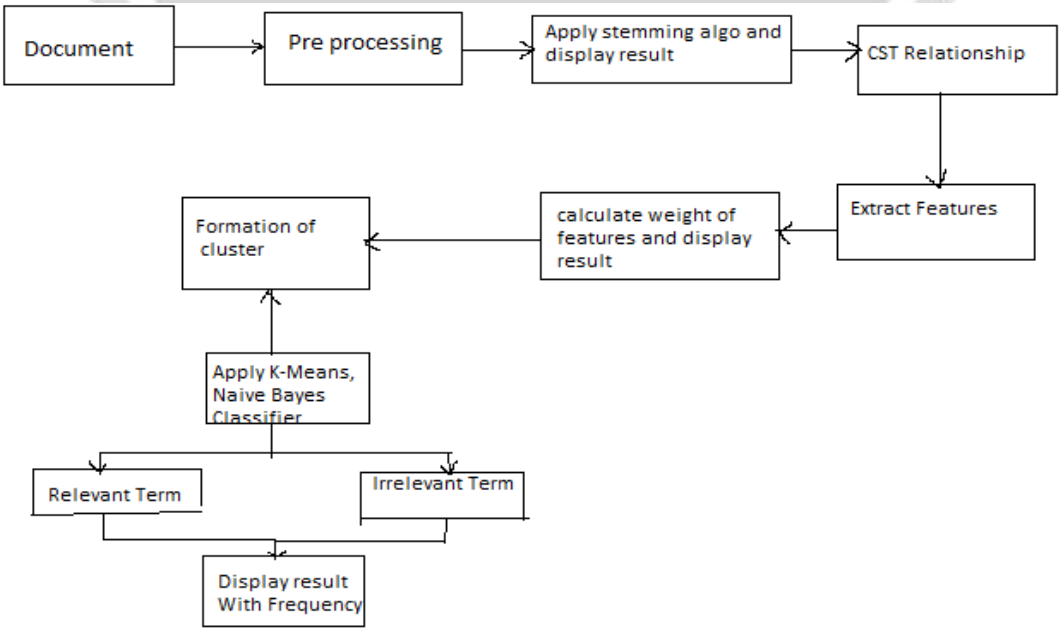

Fig -1:System Architecture

3.CST Relation- similar relation is find out between text document from which features to be extracted. length or number of thematic word by selecting single feature or all feature at a time. The basic idea has been deployed from [1] and an extension to the system is feature to find the similarity with more accuracy than the existing algorithm and Email Author Identification. 1.Document corpus- the user will choose the set of documents for checking a relevant features and similar documents within the corpus. System scans the documents uploaded by the user. 2.Preprocessing of Document- the set of text document are uploaded for the processing which removes stop words and special characters. 3.CST Relation- similar relation is find out between text document from which features to be extracted.

4.The feature extraction will be takes place according to features provided like paragraph follow title, paragraph location in document, Appearance of Sentence in paragraph, first sentence in paragraph, sentence length or number of thematic word by selecting single feature or all feature at a time. 5.weight to be calculated of extracted features according to their frequency of appearance in the document. 6.After selecting a features as per the users perspective need clusters would be form according to cluster formation relation Identity, Subsumption, Overlap or Description which may be selected single or one at a time. 7.For cluster classification K-means, Naive Bayes classifier is accustomed correct the prediction and check the performance.

4.2Pattern Taxonomy Model

There are two main stages are consider in PTM ,first one is how to extract useful phases from text documents. and second one is, how to use these discovered patterns to improve effectiveness of a knowledge discovery system. The main focus of this algorithm is deploying process, which consist of pattern discovery and term support evaluation. In this paper, we assume that all text documents are split into paragraphs. So a given document d yields a set of paragraphs PS(d). Let D be a training set of documents, which consists of a set of positive and negative documents, Let T = {t1. t2. t3. t4... tm., ..} be a set of terms (or keywords) which can be extracted from the set of positive documents.

4.3 Frequent and Closed Patterns

Frequent Patterns is one that occurs in atleast a user specific percentage of database, that percent is called support. Given a termset X in document d, X is used to denote the covering set of X for d, which includes all paragraphs dp. Its absolute support is the number of occurrences of X in PS(d). Its relative support is the fraction of the paragraphs that contain the pattern

5. RESULTS AND DISCUSSION

Reuters Corpus Volume 1, a very large data collection and TREC has provided 50 reliable assessor topics for RCV1 to filter information. RCV1 has standard collection of data including TREC 50 topics. These are sufficient for high quality of experiment. RCV1 includes 806,791 documents. Another data set Reuters-21578 has less data collection. Reuters-21578 corpus is a widely used data for text mining. Our experiment supports for RCV1. Documents in both RCV1 and Reuters-21578 are described in XML format. The result shown in below enumerate the feature extraction along with its weight for different file size. Fig.2 shows the size of the uploaded file , number of words and the count of the filter words. From the table content we predict that which document is more relevant as per the users need. Fig.3 shows the stop words and special character remove from uploaded document.

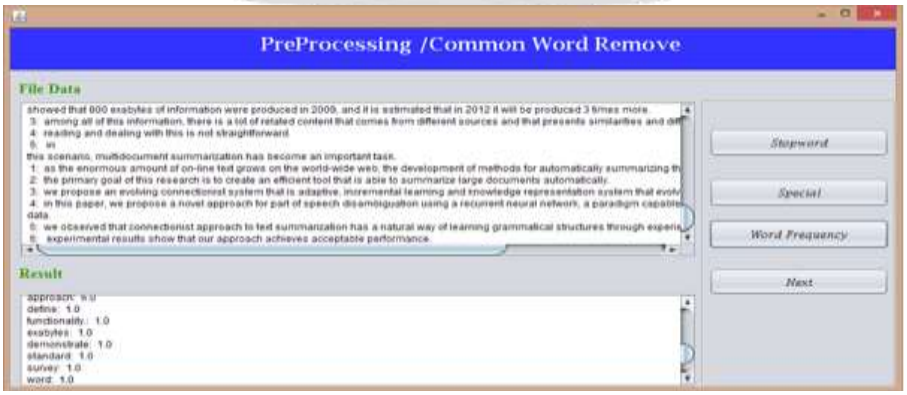

Fig -2: Remove Stop Words and Special Characters

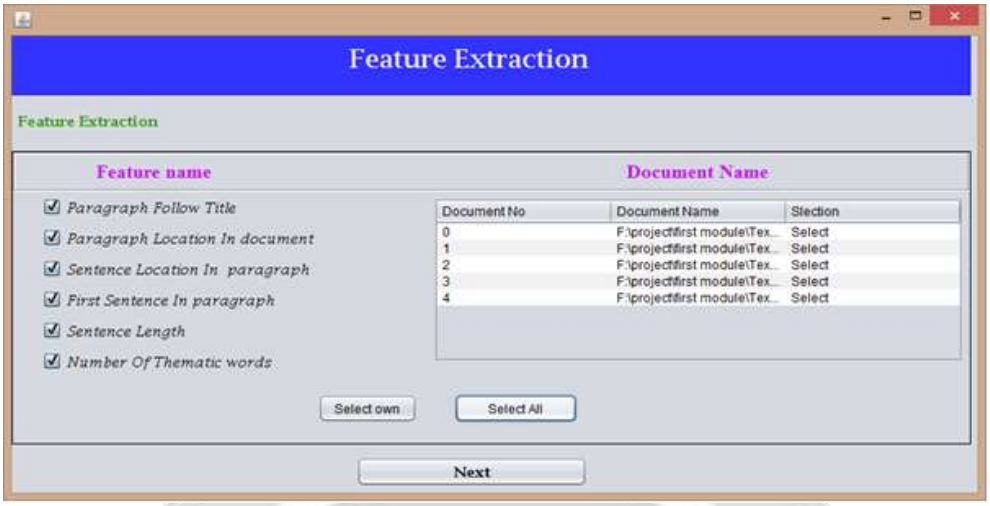

Fig-3: Feature Extraction by Selecting Pattern

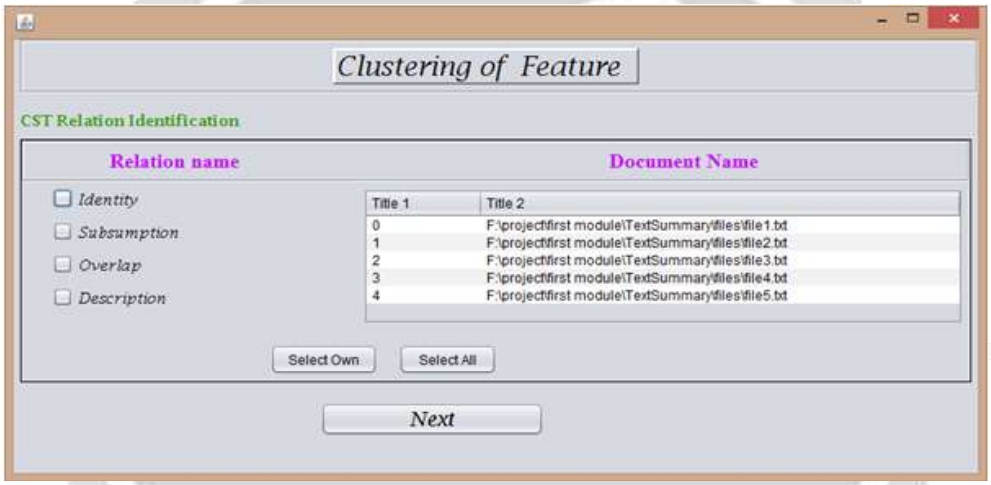

Fig-4: Cluster Formation of Extracted Features

6. CONCLUSIONS

This paper present an alternative approach for extracting the relevant feature in text document. It provides the method to filter the relevant data and classify them into different cluster according to their appearance in document. Compare with term-based model, proposed model achieve the best performance. The results also show that the term classification can be adequately approximated by the proposed feature clustering method. It provides a suitable methodology for developing efficient text mining models for relevance feature discovery.

6. REFERENCES

[1] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana, "Relevance feature discovery for text mining," VOL .27, NO.6, pp. 1656-1668, JUNE 2015.

[2] Abdulmohsen Algarni, Yuefeng Li, Xiaohui Tao," Mining Specific and General Features in Both Positive and Negative Relevance Feedback," Computer Science Discipline, Queensland University of Technology, Australia.

[3] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760-4768, 2012.

[4] R. Bekkerman and M. Gavish, "High-precision phrasebased document classification on a modern scale," in Proc. 11th ACM SIGKDD Knowl. Discovery Data Mining, 2011, pp. 231-239.

[5] T. Joachims, ”Transductive Inference for Text Classification Using Support Vector Machines,” Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 200-209, 1999.

[6] W. Lam, M.E. Ruiz, and P. Srinivasan, ”Automatic Text Categorization and Its Application to Text Retrieval,” IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[7] R. Agrawal and R. Srikant, ”Fast Algorithms for Mining Association Rules in Large Databases,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 478-499, 1994.

[8] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, ”Effective Pattern Discovery for Text Mining”,vol.24,No.1,Jan.2012.

[9]] Miss Dipti S.Charjan, Prof. Mukesh A.Pund ,” Pattern Discovery For Text Mining Using Pattern Taxonomy”, International Journal of Engineering Trends and Technology (IJETT) - Volume 4 Issue 10- October 2013.

[10] F. Debole and F. Sebastiani, ”An analysis of the relative hardness of Reuters-21578 subsets,” J. Amer. Soc. Inf. Sci. Technol., vol. 56, no. 6, pp. 584-596, 2005.

[11] J. Eisenstein, A. Ahmed, and E. P. Xing, ”Sparse additive generative models of text,” in Proc. Annu. Int. Conf. Mach. Learn., 2011, pp. 274-281.